

UC Davis

UC Davis Previously Published Works

Title

Functionally Annotating Regulatory Elements in the Equine Genome Using Histone Mark ChIP-Seq.

Permalink

<https://escholarship.org/uc/item/3g41594n>

Journal

Genes, 11(1)

ISSN

2073-4425

Authors

Kingsley, NB
Kern, Colin
Creppe, Catherine
et al.

Publication Date

2019-12-01







DOI

10.3390/genes11010003

Peer reviewed

Article

Functionally Annotating Regulatory Elements in the Equine Genome Using Histone Mark ChIP-Seq

N. B. Kingsley ^{1,2} , Colin Kern ³, Catherine Creppe ⁴, Erin N. Hales ² , Huaijun Zhou ³ ,
T. S. Kalbfleisch ⁵, James N. MacLeod ⁵, Jessica L. Petersen ⁶ , Carrie J. Finno ²  and
Rebecca R. Bellone ^{1,2,*} 

¹ Veterinary Genetics Laboratory, School of Veterinary Medicine, University of California, Davis, CA 95616, USA; nbkingsley@ucdavis.edu

² Department of Population Health and Reproduction, School of Veterinary Medicine, University of California, Davis, CA 95616, USA; enburns@ucdavis.edu (E.N.H.); cjfinno@ucdavis.edu (C.J.F.)

³ Department of Animal Science, University of California, Davis, CA 95616, USA; ckern@ucdavis.edu (C.K.); hzhou@ucdavis.edu (H.Z.)

⁴ Epigenetics Services Team, Diagenode, Liège 4102, Belgium; catherine.creppe@diagenode.com

⁵ Gluck Equine Research Center, Department of Veterinary Science, University of Kentucky, Lexington, KY 40506, USA; ted.kalbfleisch@uky.edu (T.S.K.); jnmacleod@uky.edu (J.N.M.)

⁶ Department of Animal Science, University of Nebraska, Lincoln, NE 68588, USA; jessica.petersen@unl.edu

* Correspondence: rbellone@ucdavis.edu; Tel.: +1-530-752-9299

Received: 15 November 2019; Accepted: 16 December 2019; Published: 18 December 2019



Abstract: One of the primary aims of the Functional Annotation of ANimal Genomes (FAANG) initiative is to characterize tissue-specific regulation within animal genomes. To this end, we used chromatin immunoprecipitation followed by sequencing (ChIP-Seq) to map four histone modifications (H3K4me1, H3K4me3, H3K27ac, and H3K27me3) in eight prioritized tissues collected as part of the FAANG equine biobank from two thoroughbred mares. Data were generated according to optimized experimental parameters developed during quality control testing. To ensure that we obtained sufficient ChIP and successful peak-calling, data and peak-calls were assessed using six quality metrics, replicate comparisons, and site-specific evaluations. Tissue specificity was explored by identifying binding motifs within unique active regions, and motifs were further characterized by gene ontology (GO) and protein–protein interaction analyses. The histone marks identified in this study represent some of the first resources for tissue-specific regulation within the equine genome. As such, these publicly available annotation data can be used to advance equine studies investigating health, performance, reproduction, and other traits of economic interest in the horse.

Keywords: FAANG; epigenetics; horse; genome regulation; H3K4me1; H3K4me3; H3K27ac; H3K27me3; tissue-specific; annotation

1. Introduction

In 1992, researchers discovered the first disease mutation in horses, conferring hyperkalemic periodic paralysis (HYPP) in Quarter Horses [1], yet identification of additional equine genetic diseases progressed slowly, with only nine disease-associated variants discovered prior to 2007 [2,3]. Since the release of the equine reference genome in 2007 [4], 22 additional genes were found to cause or be associated with equine diseases, yet there are at least 200 described genetic disorders for which causal variants are unknown [2,5]. While the majority of characterized equine disease variants are located within coding regions, an increasing amount of research in humans and other animal species suggests that a large number of disease mutations are harbored within regulatory elements and other functional

but non-coding regions of the genome [6–9]. Current genomic annotations for the horse have limited information about the functions of these non-coding regions, making it difficult to identify variants that alter gene regulation. While there is a high degree of conservation within coding regions across species, increasing evidence suggests that regulatory regions, especially tissue-specific elements, are not as well conserved in terms of sequence or function across species [7–10].

Using improved annotations of regulatory networks for humans and other mammals, researchers have begun to identify causal variants outside of coding regions. For example, a super-enhancer that significantly contributes to the risk of type II diabetes in humans was identified by combining a genome-wide association study (GWAS) with locations of regulatory elements and other functional regions across the genome [11]. Additionally, cattle researchers used annotations from both genomic assays on bovine tissue and homology-based methods to investigate complex traits. They found that the annotation from the homology-based method provided less relevant information than the bovine-specific annotation when combined with GWAS pertaining to milk production [8]. Along with the increasing number of regulatory GWAS, researchers are also expanding efforts to look at large-scale changes in the epigenome, such as associations between genome-wide changes in active regulatory elements and autism spectrum disorder [12]. With the increasing focus on the importance of regulatory elements in the pathogenesis of many diseases, it is clear that annotations of the equine genome need to encompass both coding and non-coding functional elements in order to understand complex genetic diseases and other traits affecting horses and the equine industry.

Similar to the Encyclopedia of DNA Elements (ENCODE), the Functional Annotation of ANimal Genomes (FAANG) initiative was established to improve the reference annotation of animal species, including characterization of genomic regulatory regions [13–15]. The link between chromatin modifications and regulatory regions, such as enhancers and promoters, has been well established for several decades [16]. Four histone tail modifications, also known as histone marks, were selected by the FAANG consortium to characterize promoters (H3K4me3) and enhancers (H3K4me1), as well as distinguish between active (H3K27ac) and inactive (H3K27me3) genomic elements [15].

Histone tail modifications were first hypothesized to affect genomic regulation in 1964 [17]. Since that time, associations between histone marks and regulatory elements have continued to expand. However, it is still unclear if the associations are the result of underlying functional roles in all cases or if the histone marks are the by-products of regulatory activities [18]. There is evidence that monomethylation of H3K4 recruits chromatin-modifying proteins to enhancer regions leading to enhancer priming and recognition [19]. Conversely, H3K4me3 is associated with promoters and the transcription start site, but some suggest that the mark may actually be a signature left by frequent transcriptional activity to create cellular memory rather than acting as a pioneer factor [18]. In conjunction with the H3K4 modifications, acetylation of H3K27 is found at active enhancers and promoters [20]. Trimethylation of the same residue, however, is strongly associated with facultative heterochromatin, leading to inactive regulatory elements and silenced genes [20]. In fact, these two H3K27 marks are thought to be antagonistic to one another, such that acetylation may actually prevent Polycomb silencing [21]. Acetylation of H3K27 also decreases the overall positive charge of the histone proteins, leading to fewer interactions with DNA and more open chromatin structure [22]. Since H3K27ac has been strongly associated with active regulatory elements across the genome, presence of this mark can be used to distinguish between active and poised regulatory elements [23]. While research into the functional roles of histone tail modifications continues, the associations between these four marks and patterns of regulatory elements are well established. For more than a decade, chromatin immunoprecipitation (ChIP) assays have remained the primary method for identifying genomic sites enriched with histone marks to conduct large-scale regulatory mapping [24,25].

As part of the international FAANG collaboration, we performed ChIP sequencing (ChIP-Seq) on four major histone modifications (H3K4me1, H3K4me3, H3K27ac, and H3K27me3) in eight equine tissues (adipose, brain (parietal cortex), heart, lamina, liver, lung, (skeletal) muscle, and ovary). These tissues were collected as part of the FAANG equine biobank, which includes samples from

80 tissues, six fluids, and two cell lines that were collected from two healthy adult thoroughbred mares [26]. The eight tissues evaluated in this study were prioritized for thorough investigation based on (1) continuity with other FAANG species to enable across species comparisons and/or (2) the primary needs of the equine community in terms of health, performance, and reproduction.

2. Materials and Methods

Due to the limited number of previous histone ChIP-Seq experiments across equine tissues, we first performed quality control (QC) testing to determine appropriate experimental parameters for each tissue and mark. Sequencing, rather than targeted qPCR, was used to assess the quality of the QC data due to the limited knowledge of appropriate tissue-specific control genes in the horse. All ChIP-Seq experiments, including QC, were conducted by Diagenode ChIP-Seq Profiling Service (Diagenode, Cat# G02010000, Liège, Belgium), and a complete summary of the final protocols used for all tissues can be accessed at <http://ftp.faang.ebi.ac.uk/ftp/protocols/assays/>. Quality control involved chromatin extraction, ChIP, library preparation, and sequencing of one training sample from each of the eight tissues investigated. Quality control training samples were obtained from previously banked tissues collected at the University of California, Davis. Bioinformatic analysis was performed on the QC data in order to calculate library complexity and ChIP enrichment metrics and to call peaks to evaluate the genomic distribution of detected marks. Library complexity metrics included non-redundant fraction (NRF) and PCR bottleneck coefficients 1 and 2 (PBC1 and PBC2). Metrics for ChIP enrichment included normalized strand cross-correlation coefficient (NSC), relative strand cross-correlation coefficient (RSC), and Jensen–Shannon distance (JSD). Non-redundant fraction, PBC1 and 2, NSC, and RSC are all standardized metrics of the ENCODE project [27] and were compared against ENCODE standards to determine the efficacy of the ChIP protocols. Jensen–Shannon distance is a common statistic used to compare two distributions that can be applied to ChIP datasets using deepTools version 2.4.3 [28], and a threshold was determined by agreement among FAANG collaborators. Tissues of interest (TOI) used in the final experiments that generated combined peak-calls were collected from two thoroughbred mares (referred to in this manuscript as ECA_UCD_AH1 for SAMEA104728862 and ECA_UCD_AH2 for SAMEA104728877) as part of the FAANG equine biobank [26], and all protocols for this work were approved by the University of California, Davis Institutional Animal Care and Use Committee (Protocol #19037). Laboratory procedures that varied based on tissue are summarized in Table 1.

Table 1. Optimized ChIP-Seq experimental procedures for each tissue.

Parameter	Adipose	Brain	Heart	Lamina	Liver	Lung	Muscle	Ovary
Starting Tissue (mg)	220	90	105	100	40	40	100	85
Homogenization Time (min)	8	5	5	9	n/a	5	5	5
Duration Fixed (min)	10	9	9	9	9	9	9	9
Fixation Temp. (°C)	37	23	23	23	23	23	23	23
Shearing Volume (uL)	400	1500	1800	1800	1500	1500	1800	1800
Shearing Cycles	5 × 8 ¹	10	13	10	12	10	12	10
Chromatin per IP (ng)	700	300	500	700	450	800	260	1200

¹ Samples were sonicated for five sets of eight cycles.

2.1. Chromatin Extraction

Chromatin was extracted from adipose using the True MicroChIP Kit (Diagenode, Cat# C01010130, Liège, Belgium) and from the other seven tissues following the iDeal ChIP-Seq Kit for histones (Diagenode, Cat# C01010059, Liège, Belgium) with the modifications or specifications described in this paper. The starting amount varied depending on tissue, such that those with extensive extracellular matrices and/or low ratio of nuclei to cellular matter required larger amounts of starting material compared to those tissues that homogenized easily. Samples were homogenized either by douncing (liver) or grinding with the Tissue Lyser II (Qiagen, Germany) at 25 strokes/minute for a length of time that varied by tissue (Table 1).

In order to reach the desired fragment length (approximately 200 bp), chromatin was sheared with the Bioruptor Pico (Diagenode, Cat# B01060001, Liège, Belgium) combined with the Bioruptor® Water cooler for 8–12 cycles of 30 s with 30 s rest between cycles. The temperature was maintained at 10 °C for adipose and 4 °C for all other tissues during shearing. The number of cycles varied based on tissue (Table 1), and the chromatin quality was assessed using the Fragment Analyzer (Aligent, Santa Clara, CA, USA).

2.2. Immunoprecipitation

Immunoprecipitation (IP) of the four histone marks, along with a negative IP control (IgG), was performed on tissue-specific amounts of chromatin using the IP-Star Compact Automated System (Diagenode, Cat# B03000002, Liège, Belgium). The antibodies used were all previously validated by Diagenode, and antibody concentrations were determined during QC for every tissue and histone mark combination (Table S1). An aliquot of chromatin from each tissue was set aside for the input to characterize sequencing background and identify true ChIP enrichment.

2.3. Sequencing

Libraries were prepared using the IP-Star® Compact Automated System (Diagenode, Cat# B03000002, Liège, Belgium) and the MicroPlex Library Preparation Kit v2 (Diagenode, Cat# C05010013, Liège, Belgium) for the input and four ChIPs per tissue. Libraries were amplified prior to sequencing for at least 10 cycles, and additional cycles were performed as needed to reach a concentration of 3–10 nM. Using Agencourt AMPure XP (Beckman Coulter, Brea, CA, USA), libraries were purified and fragments were size-selected for approximately 200 bp. Libraries were sequenced as 50 bp, single-end reads on the HiSeq 4000 platform (Illumina, San Diego, CA, USA) to generate approximately 55–80 M raw reads for H3K27me3 and 30–50 M raw reads for the other marks and inputs.

2.4. Data Processing

A complete summary of the bioinformatic workflow can be accessed at <ftp://ftp.faang.ebi.ac.uk/ftp/protocols/analysis/>, and bioinformatic parameters that varied by mark are summarized in Table 2. Reads were trimmed using Trim-Galore version 0.4.0 [29,30] under the default parameters and aligned to EquCab3.0 [31] with BWA-MEM version 0.7.9a [32], such that split hits were marked as secondary alignments. Alignments were converted to BAM file format, processed, and filtered using SAMtools version 1.9 [33]. For strict quality filtering, reads were removed if they did not map, had secondary alignments, failed platform/vendor quality tests, were identified as optical duplicates, or had an alignment quality score lower than 30. PCR duplicates were marked with PicardTools version 2.7.1 [34] and then removed with SAMtools. For peak-calling, MACS2 version 2.1.1.20160309 [35] was used to call peaks for all marks, and SICERpy version 0.1.1 [36], which is a wrapper for executing SICER [37], was also used to call peaks for the broad mark, H3K27me3. Combining peak-calls involved identifying overlapping regions of enrichment in both biological replicates where at least one replicate was significantly enriched based on a set of significance thresholds that varied by mark (Table 2). Additionally, enrichment tracks (bigWig files) were generated using deepTools version 2.4.3, which subtracted background characterized by the input and then combined enrichment from both replicates.

Table 2. Software parameters used to analyze ChIP-Seq data for each histone mark.

Software	Parameter	H3K4me1	H3K4me3	H3K27ac	H3K27me3
MACS2	Filtering	strict	strict	strict	strict
	Size	narrow ¹	narrow	narrow	broad
	Size Flag	none	none	none	–broad
	Model	–fix-bimodal	–fix-bimodal	–fix-bimodal	–fix-bimodal
	Genome Size	2,409,143,234	2,409,143,234	2,409,143,234	2,409,143,234
both	Fragment Size	200	200	200	200
	FDR	0.05	0.01	0.01	0.1
SICERpy ²	Gap Size	n/a	n/a	n/a	4
	Window Size	n/a	n/a	n/a	200
	Genome Fraction	n/a	n/a	n/a	0.63

¹ Mark was treated as broader than other narrow marks due to being categorized previously as broad by ENCODE [19]. ² SICERpy was only used to call peaks for the broad mark, H3K27me3.

2.5. Data Analysis

As with the QC data, the datasets from the eight TOI were assessed by calculating library complexity and ChIP enrichment metrics, as well as evaluating the genomic distribution of detected marks. Identity between peaks called for ECA_UCD_AH1 and ECA_UCD_AH2 were compared to assess the similarity of the biological replicates using the Jaccard Index [38], also known as the Jaccard Similarity Coefficient, from BEDtools version 2.27.1 [39]. Unique peaks were defined as a peak for a given mark that does not have any overlap with peaks from the same mark in the other prioritized tissues. BEDtools version 2.27.1 was used to identify unique peaks, as well as calculate the percent of the genome covered by peaks. Graphs were generated using ggplot2 with R software version 3.4.3 [40,41]. To characterize the average peak topology in relation to gene annotations and calculate normalized enrichment patterns, we used deepTools version 2.4.3. RNA-Seq data from the two FAANG replicates allowed for site-specific validation of the histone peaks (ERR2584116, ERR2584168, ERR2584153, ERR2584205, ERR2584194, ERR2584142, ERR2584135, ERR2584187, ERR2584195, ERR2584143, ERR2584197, ERR2584145, ERR2584144, ERR2584196, ERR2584152, and ERR2584204). Analysis of Motif Enrichment (AME) from the MEME Suite version 5.0.5 [42] was used to identify known transcription factor binding motifs within peaks based on the JASPAR 2018 vertebrate database [43], and Biological Process Gene Ontologies (GOs) from Swiss Prot were used to perform a GO term analysis [44]. Novel motifs were characterized with DREME and MEME, and each of the novel motifs was manually investigated to identify additional known motifs that were not included in the JASPAR database. The Integrated Genome Viewer [45] was used to visualize peak-calls in conjunction with the Ensembl annotation (release 95) for the EquCab3.0 reference assembly [46]. String version 11.0 was used to perform a protein–protein interaction analysis on the transcription factors implicated in each tissue based on the enriched motifs identified in the tissue-specific active enhancer elements [47].

3. Results

Quality control testing was performed to determine tissue-specific laboratory parameters such as antibody concentration and shearing time (Table 1 and Table S1) by comparing the quality metrics and peak-calls to ENCODE and FAANG standards (Table S2). The raw and processed data are available on <https://data.fang.org/home> under the study accession PRJEB35307. The processed files include read alignments to EquCab3.0 and peak-calls for each biological replicate, as well as the combined peak-calls.

3.1. Assessing Data Quality

Using the Jaccard Index to compare replicates for all of the marks, we found the highest identity between the two replicates of the same tissue (Figure 1), with the exception of the brain replicates for H3K4me1 and the ovary replicates for H3K27me3. For the brain replicates, ECA_UCD_AH2 had

65,327 peaks compared to 143,328 peaks for ECA_UCD_AH1 (Table 3). In addition to a lower peak number, two of the library complexity quality scores for the ECA_UCD_AH2 brain replicate were lower than the ENCODE quality thresholds (Table 4). In terms of enrichment, however, the cross-correlation metrics for this sample were both above the established thresholds and indicate that the data had sufficient ChIP enrichment. Indeed, we were able to call 95,918 combined peaks, which is consistent with the range of H3K4me1 values from the other tissues. For the ovary replicates, two of the three library complexity metrics were below ENCODE standards, yet all of the ChIP enrichment metrics were consistent with those for the broad mark in other tissues, indicating that we had sufficient ChIP enrichment despite lower library complexity. When comparing SICER and MACS2 peak-calls for the broad mark, the number of combined peaks (8479 and 40,825, respectively) and the percentage of the genome covered (2.1% and 1.2%, respectively) for ovary were all consistent with the same measures for the H3K27me3 peaks from the other tissues evaluated in this study.

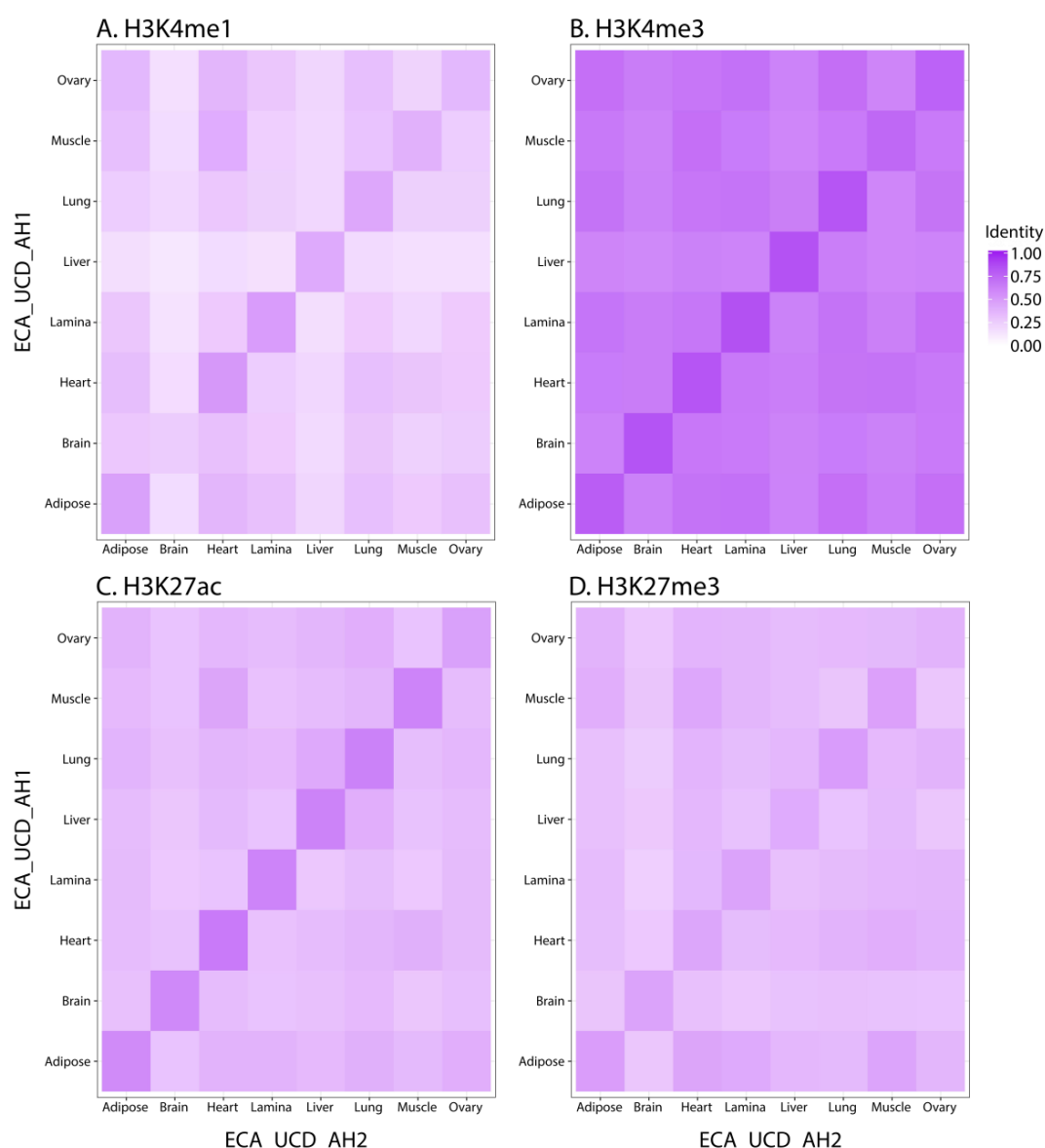


Figure 1. Jaccard Index to measure similarity of peaks called for each histone mark and tissue between biological replicates, ECA_UCD_AH1 and ECA_UCD_AH2. Each panel is a heatmap displaying the Jaccard Index for pairwise comparisons of tissues between replicates. Darker purple indicates that there are more peaks that are shared by the two tissues. (A) H3K4me1, (B) H3K4me3, (C) H3K27ac, (D) H3K27me3 called by SICER.

Table 3. Summary of peak number and percent of the genome covered for each mark and tissue.

Mark	Tissue	Software	Combined Peak Number	Percent Genome Covered	AH1 Peak Number	AH2 Peak Number
H3K4me1	Adipose	MACS2	107,318	5.1	130,242	157,497
	Brain	MACS2	95,918	3.1	143,328	65,327
	Heart	MACS2	121,663	4.9	137,385	155,881
	Lamina	MACS2	114,708	4.2	137,575	124,150
	Liver	MACS2	116,760	3.6	97,863	135,122
	Lung	MACS2	92,972	2.9	90,687	109,001
	Muscle	MACS2	95,816	3.7	137,322	100,999
	Ovary	MACS2	102,986	4.3	166,303	133,209
H3K4me3	Adipose	MACS2	26,905	1.7	26,286	29,121
	Brain	MACS2	27,101	1.6	25,473	28,028
	Heart	MACS2	26,475	1.4	24,101	27,985
	Lamina	MACS2	29,380	1.6	29,023	19,742
	Liver	MACS2	28,498	1.5	28,204	28,222
	Lung	MACS2	28,546	1.6	30,048	27,779
	Muscle	MACS2	28,110	1.6	30,428	25,123
	Ovary	MACS2	28,378	1.7	30,522	29,192
H3K27ac	Adipose	MACS2	79,620	3.3	75,823	99,249
	Brain	MACS2	78,823	3.2	89,445	73,795
	Heart	MACS2	68,728	2.9	71,462	7192
	Lamina	MACS2	82,394	2.9	91,345	78,953
	Liver	MACS2	87,589	3.1	84,814	96,238
	Lung	MACS2	69,054	2.9	69,621	75,299
	Muscle	MACS2	76,495	2.9	78,047	86,524
	Ovary	MACS2	64,318	3.3	94,817	82,318
H3K27me3	Adipose	MACS2	25,183	0.6	8948	29,906
	Brain	MACS2	24,243	0.6	16,055	23,411
	Heart	MACS2	68,113	1.8	31,455	88,818
	Lamina	MACS2	37,366	0.8	31,839	28,508
	Liver	MACS2	63,874	1.3	93,423	23,888
	Lung	MACS2	30,191	0.7	32,385	18,124
	Muscle	MACS2	42,610	0.9	39,076	29,579
	Ovary	MACS2	40,825	1.2	43,036	33,220
H3K27me3	Adipose	SICER	8167	4.9	13,540	14,571
	Brain	SICER	7860	3.4	11,386	13,603
	Heart	SICER	9032	3.3	12,192	18,903
	Lamina	SICER	7072	3.8	11,933	11,694
	Liver	SICER	11,430	3.7	22,270	16,099
	Lung	SICER	7863	2.6	12,668	11,715
	Muscle	SICER	8437	4.6	17,073	10,987
	Ovary	SICER	7083	3.0	14,731	11,124

While demonstrating more variation than the other marks, H3K27me3 had peaks called by SICER that were more similar between replicates of the same tissue compared to the H3K27me3 peaks called by MACS2 software (Figure S1). Peak-calls from the two biological replicates were combined by identifying regions of overlapping enrichment in which at least one replicate had significant enrichment based on a q-value that differed by mark (Table 2). For H3K4me1, H3K4me3, and H3K27ac, the number of combined peaks ranged from 93–121 K, 26–29 K, and 64–88 K, respectively. The number of combined peaks called for the broad mark was lower than the three activating marks for both MACS2 and SICER (24–68 K and 7–11 K, respectively). Although the combined peak numbers for the MACS2-H3K27me3 datasets were more similar to the ENCODE equivalent than the number of peaks called by SICER (Table S3), the SICER-H3K27me3 combined peaks covered a larger portion of the genome, similar to that for the other marks (Table 3). Files for H3K27me3 peaks called by MACS2 and SICER are both publicly available for every tissue.

As a proof-of-principle, we investigated a small number of regions near well-characterized genes to compare the histone peaks with RNA-seq data generated from the same tissues. We found consistent activating marks across all tissues for a widely expressed gene, *ACTB* (Figure 2A). Conversely, liver was the only tissue enriched for a set of active histone marks near the transcription start site (TSS) of the liver-specific gene *CYP2E1* (Figure 2B) [48].

Table 4. Quality metrics for assessing library complexity and ChIP enrichment. Thresholds for NRF, PBC1, PBC2, NSC, and RSC represent those developed by the Encyclopedia of DNA Elements (ENCODE) [27]. JSD threshold was established among members of the Functional Annotation of ANimal Genomes (FAANG) consortium.

Mark	Tissue	Rep	NRF	PBC1	PBC2	NSC	RSC	JSD
	Thresholds		(>0.5)	(>0.5)	(>1)	(>1.05)	(>0.8)	(>0.05)
H3K4me1	Adipose	AH2	0.677	0.673	3.017	1.068	1.249	0.281
	Adipose	AH1	0.621	0.617	2.595	1.067	1.147	0.239
	Brain	AH2	0.435	0.443	1.908	1.055	1.243	0.186
	Brain	AH1	0.754	0.756	4.128	1.074	1.275	0.228
	Heart	AH2	0.708	0.708	3.444	1.086	1.790	0.321
	Heart	AH1	0.497	0.496	2.023	1.071	1.657	0.259
	Lamina	AH2	0.606	0.606	2.551	1.093	1.628	0.281
	Lamina	AH1	0.561	0.562	2.311	1.088	1.809	0.283
	Liver	AH2	0.760	0.762	4.226	1.097	1.240	0.226
	Liver	AH1	0.838	0.842	6.462	1.117	1.289	0.252
	Lung	AH2	0.736	0.736	3.796	1.079	1.123	0.199
	Lung	AH1	0.667	0.665	2.980	1.069	1.063	0.178
	Muscle	AH2	0.706	0.706	3.418	1.077	1.030	0.210
	Muscle	AH1	0.576	0.573	2.338	1.084	1.200	0.259
	Ovary	AH2	0.712	0.712	3.488	1.077	1.265	0.245
	Ovary	AH1	0.692	0.691	3.240	1.085	2.117	0.313
H3K4me3	Adipose	AH2	0.595	0.604	2.581	1.322	1.391	0.382
	Adipose	AH1	0.559	0.571	2.389	1.313	1.501	0.354
	Brain	AH2	0.497	0.515	2.167	1.366	1.198	0.516
	Brain	AH1	0.333	0.362	1.813	1.360	1.249	0.528
	Heart	AH2	0.410	0.435	1.905	1.467	1.364	0.540
	Heart	AH1	0.337	0.374	1.857	1.399	1.639	0.548
	Lamina	AH2	0.529	0.551	2.345	1.384	1.188	0.467
	Lamina	AH1	0.571	0.594	2.606	1.380	1.289	0.465
	Liver	AH2	0.452	0.471	1.996	1.407	1.196	0.517
	Liver	AH1	0.421	0.444	1.926	1.385	1.282	0.537
	Lung	AH2	0.610	0.628	2.813	1.354	1.154	0.387
	Lung	AH1	0.580	0.600	2.634	1.344	1.117	0.452
	Muscle	AH2	0.240	0.277	1.818	1.340	1.354	0.441
	Muscle	AH1	0.559	0.567	2.350	1.350	1.164	0.448
	Ovary	AH2	0.633	0.646	2.926	1.315	1.191	0.428
	Ovary	AH1	0.603	0.622	2.779	1.335	1.220	0.439
H3K27ac	Adipose	AH2	0.678	0.677	3.087	1.223	1.605	0.313
	Adipose	AH1	0.537	0.532	2.129	1.250	1.800	0.333
	Brain	AH2	0.495	0.493	2.001	1.202	1.320	0.310
	Brain	AH1	0.655	0.657	2.939	1.200	1.341	0.326
	Heart	AH2	0.493	0.489	1.970	1.316	2.193	0.402
	Heart	AH1	0.573	0.573	2.361	1.331	1.856	0.376
	Lamina	AH2	0.597	0.596	2.486	1.296	1.655	0.351
	Lamina	AH1	0.657	0.662	3.006	1.304	1.711	0.345
	Liver	AH2	0.719	0.722	3.651	1.258	1.225	0.347
	Liver	AH1	0.721	0.724	3.674	1.242	1.237	0.298
	Lung	AH2	0.500	0.499	2.008	1.241	1.290	0.327
	Lung	AH1	0.654	0.658	2.956	1.208	1.281	0.299
	Muscle	AH2	0.605	0.604	2.524	1.291	1.306	0.335
	Muscle	AH1	0.510	0.511	2.072	1.285	1.335	0.381
	Ovary	AH2	0.733	0.736	3.816	1.254	1.309	0.374
	Ovary	AH1	0.678	0.678	3.112	1.224	1.461	0.391

Table 4. Cont.

Mark	Tissue	Rep	NRF	PBC1	PBC2	NSC	RSC	JSD
	Thresholds		(>0.5)	(>0.5)	(>1)	(>1.05)	(>0.8)	(>0.05)
H3K27me3	Adipose	AH2	0.646	0.641	2.751	1.057	0.659	0.101
	Adipose	AH1	0.650	0.647	2.809	1.053	0.592	0.067
	Brain	AH2	0.511	0.510	2.077	1.060	0.400	0.101
	Brain	AH1	0.616	0.614	2.587	1.067	0.477	0.088
	Heart	AH2	0.407	0.414	1.834	1.070	0.595	0.106
	Heart	AH1	0.287	0.315	1.778	1.090	0.649	0.102
	Lamina	AH2	0.459	0.460	1.919	1.069	0.656	0.071
	Lamina	AH1	0.429	0.436	1.885	1.076	0.732	0.065
	Liver	AH2	0.545	0.537	2.140	1.076	0.648	0.093
	Liver	AH1	0.454	0.451	1.871	1.084	0.661	0.123
	Lung	AH2	0.619	0.615	2.575	1.072	0.617	0.072
	Lung	AH1	0.550	0.545	2.199	1.084	0.671	0.088
	Muscle	AH2	0.534	0.526	2.098	1.070	0.597	0.070
	Muscle	AH1	0.476	0.472	1.914	1.079	0.689	0.103
	Ovary	AH2	0.524	0.520	2.103	1.071	0.587	0.066
	Ovary	AH1	0.495	0.489	1.970	1.077	0.688	0.101

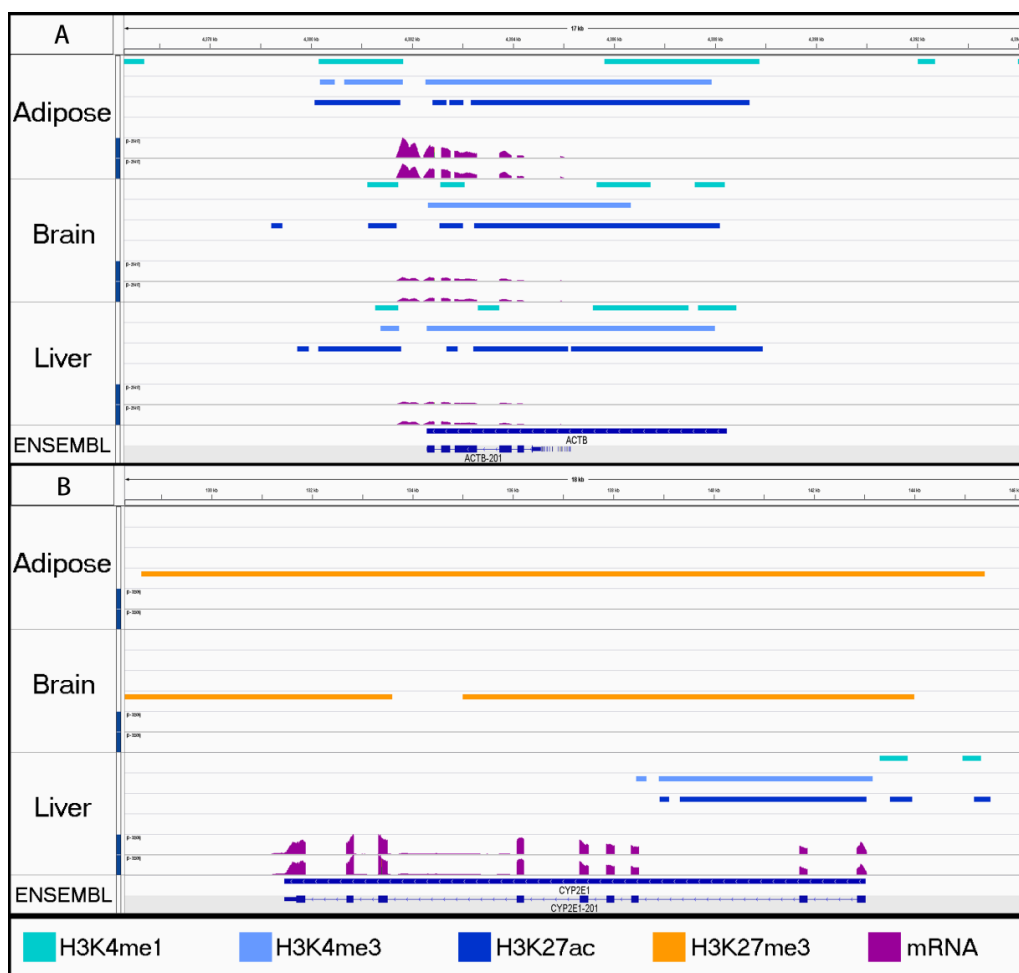


Figure 2. Proof-of-principle investigating house-keeping gene, *ACTB*, and liver-specific gene, *CYP2E1*, for appropriate regulatory elements. For adipose, brain, and liver tissue, combined peaks are displayed for H3K4me1 (aqua), H3K4me3 (light blue), H3K27ac (dark blue), H3K27me3 from SICER (orange), and mRNA expression (purple). (A) *ACTB* is a housekeeping gene that is commonly expressed for many tissues. (B) *CYP2E1* is a liver enzyme, which displays tissue-specific expression. Note the presence of the H3K27me3 repressive mark (orange) within adipose and brain samples.

3.2. Characterizing Tissue-Specific Features

Brain tissue had the highest percentage of unique peaks, defined as peaks that were only found in that tissue, for H3K27ac (31%) and H3K27me3 (20%), while liver had the highest percentage of unique peaks for H3K4me1 (32%) and H3K4me3 (16%), along with the second highest for H3K27ac (26%) and H3K27me3 (14%) (Figure 3). Lamina tissue also had a high percentage of unique peaks for the three activating marks with 24, 10, and 26% for H3K4me1, H3K4me3, and H3K27me3, respectively.

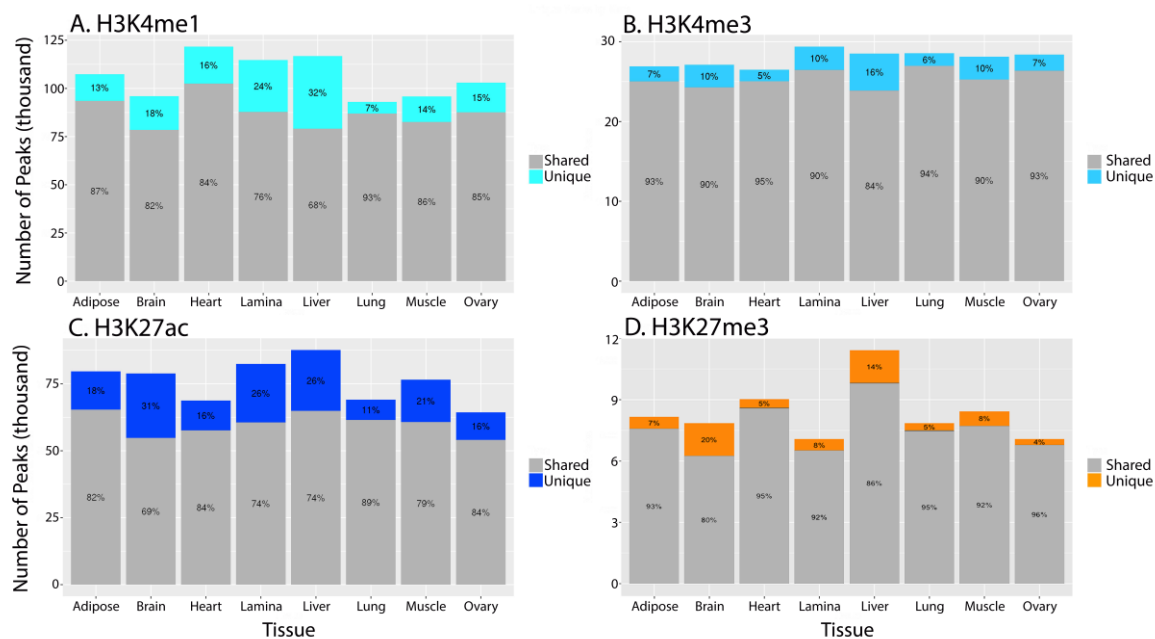


Figure 3. Tissue-specific peaks for each histone mark. Grey area indicates the number of peaks for a particular histone tail modification that are shared between at least two tissues, while the color region of each bar indicates the number of tissue-specific peaks. Percentage values are also assigned to the two segments of each bar to indicate the proportion of shared and unique peaks. (A) H3K4me1, (B) H3K4me3, (C) H3K27ac, (D) H3K27me3 from SICER.

In addition to characterized genes, we also investigated a small number of genomic regions with putative tissue-specific functions in liver and muscle. For liver, a potential tissue-specific regulatory element was identified in the 59th intron of *PKHD1* (Ensembl Transcript ID: ENSECAT00000024985.1; Figure 4), a gene which has been previously associated with liver fibrosis [49]. Similarly, when considering a genomic region associated with racing ability [50], peaks for H3K4me3 in both muscle tissues were discovered at the start of a predicted lncRNA from Ensembl genebuild [51], indicating that this uncharacterized gene may be particularly informative for the function of contractile tissue (Figure 5).

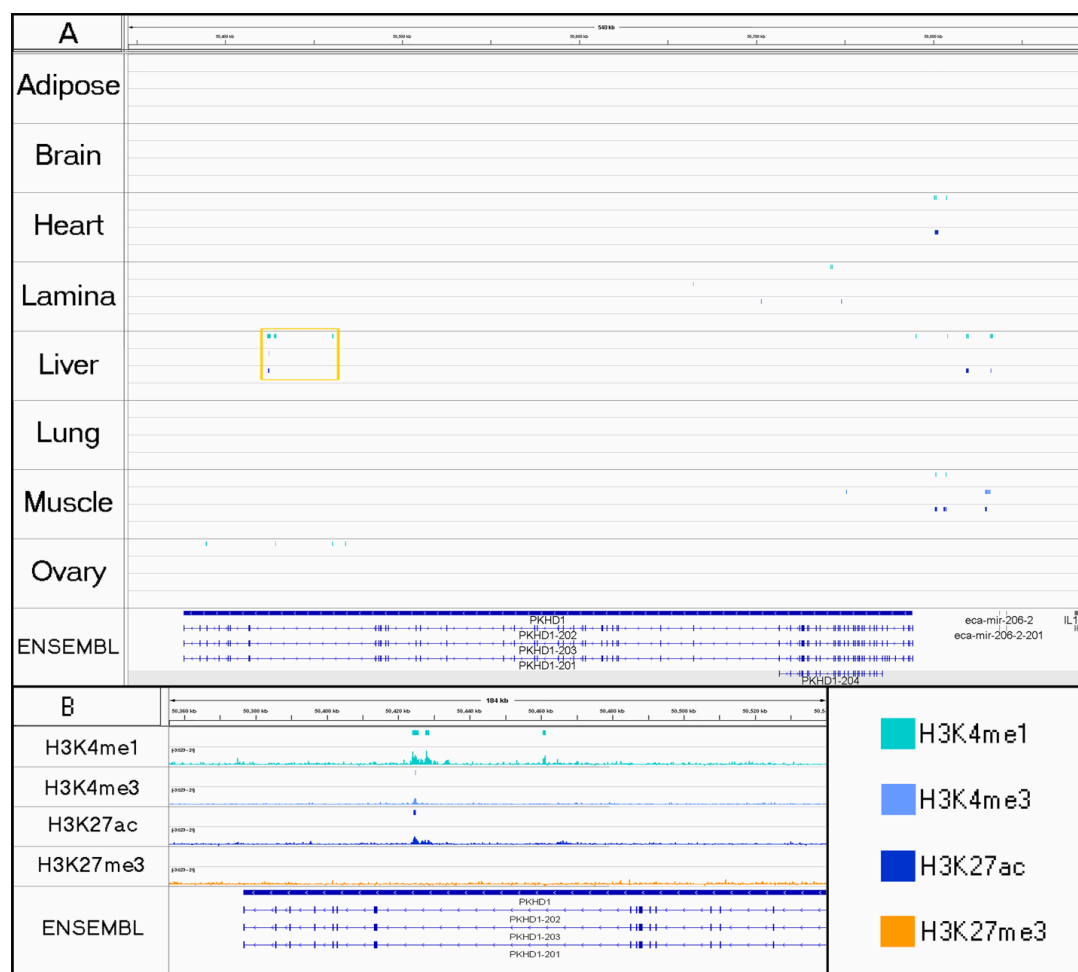


Figure 4. Evidence of a tissue-specific regulatory element found in liver tissue. For each tissue, peaks are displayed for H3K4me1 (aqua), H3K4me3 (light blue), H3K27ac (dark blue), and H3K27me3 from SICER (orange). **(A)** Gold box highlights liver-specific active marks in the 59th intron of an annotated gene, *PKHD1* (Ensembl Transcript ID: ENSECAT00000024985.1), which is transcribed from the antisense strand. H3K4me1 marks were also detected in ovary tissue at the end of the gene, but they do not indicate the presence of an active enhancer without co-occurrence of H3K27ac. **(B)** Enrichment profiles (BigWig) were visualized below the corresponding peak tracks for the region highlighted by the gold box in A.

Across all tissues, H3K4me1 was enriched around the TSS, with a decrease in enrichment at the actual annotated start site that created a bimodal distribution across the average gene body (Figure 6A). Additionally, H3K4me1 maintained a moderate level of enrichment throughout the gene body, as well as 3 Kb up- and downstream, as expected for distal and proximal enhancer elements. Alternatively, H3K4me3 (Figure 6B) and H3K27ac (Figure 6C) had peaks of enrichment at the TSS, although H3K27ac also showed enrichment, to a lesser extent, just upstream of the TSS. H3K27me3 had lower enrichment than the other three marks overall, but the average enrichment was essentially constant throughout the gene body, as well as 3 Kb up- and downstream (Figure 6D). While patterns of enrichment for each mark were highly consistent between tissues, the enrichment of H3K27me3 for lamina was lower at the TSS compared to the other tissues.

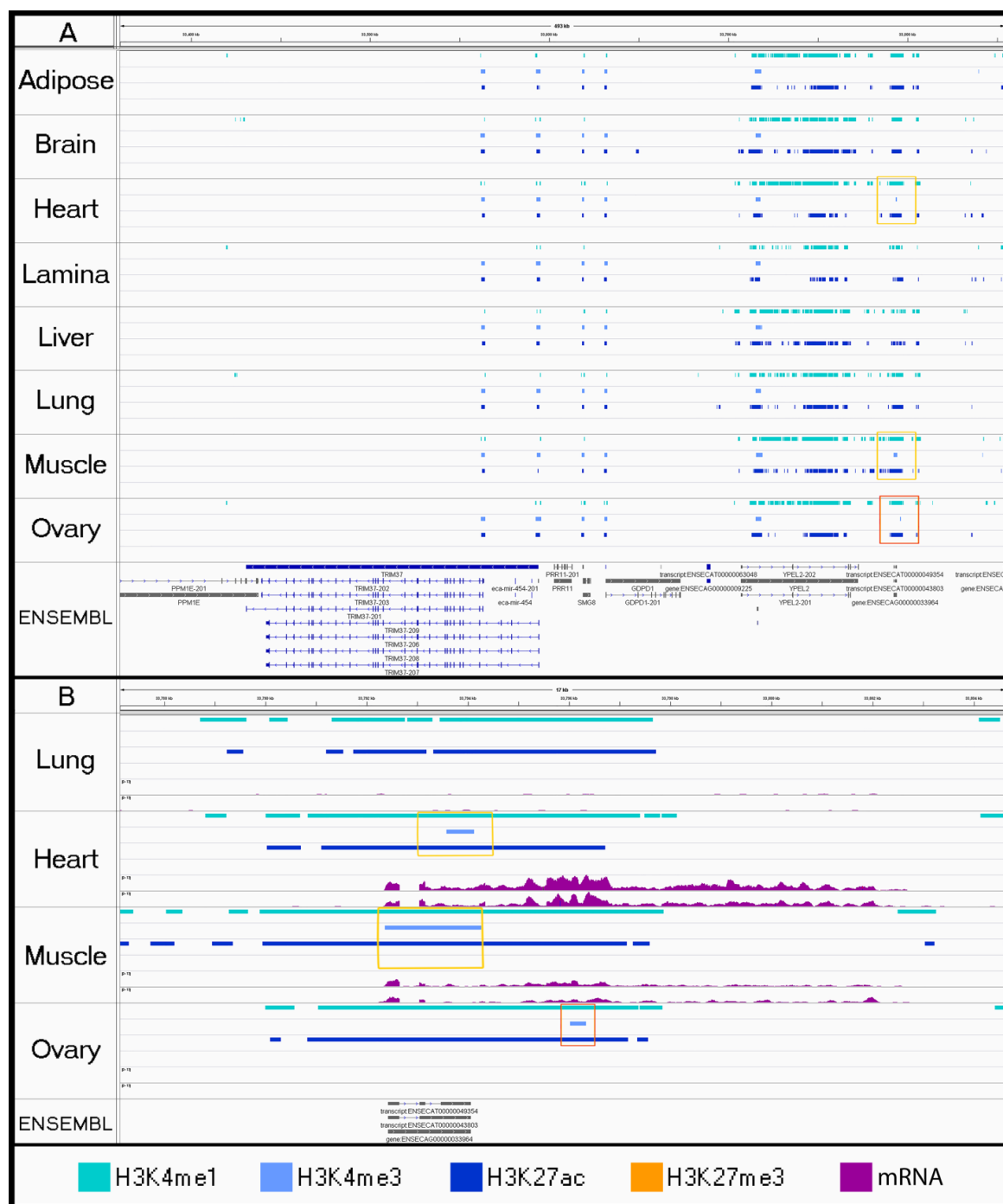


Figure 5. Visualizing tissue-specific peak-calls using the Integrated Genome Viewer. For each tissue, peaks are displayed for H3K4me1 (aqua), H3K4me3 (light blue), H3K27ac (dark blue), and H3K27me3 from SICER (orange). Gold boxes highlight active marks associated with promoters (H3K4me3) in both muscle tissues (skeletal and cardiac) for an unannotated ncRNA, *LOC111775680* (*ENSECAT00000049354*), and red box highlights an H3K4me3 peak specific to ovary tissue. **(B)** Zoomed in view of **(A)** for relevant tissues with RNA expression shown in purple. While ovary also appears to have a peak in H3K4me3 nearby, it does not have expression of the lncRNA based on mRNA expression from these tissue samples.

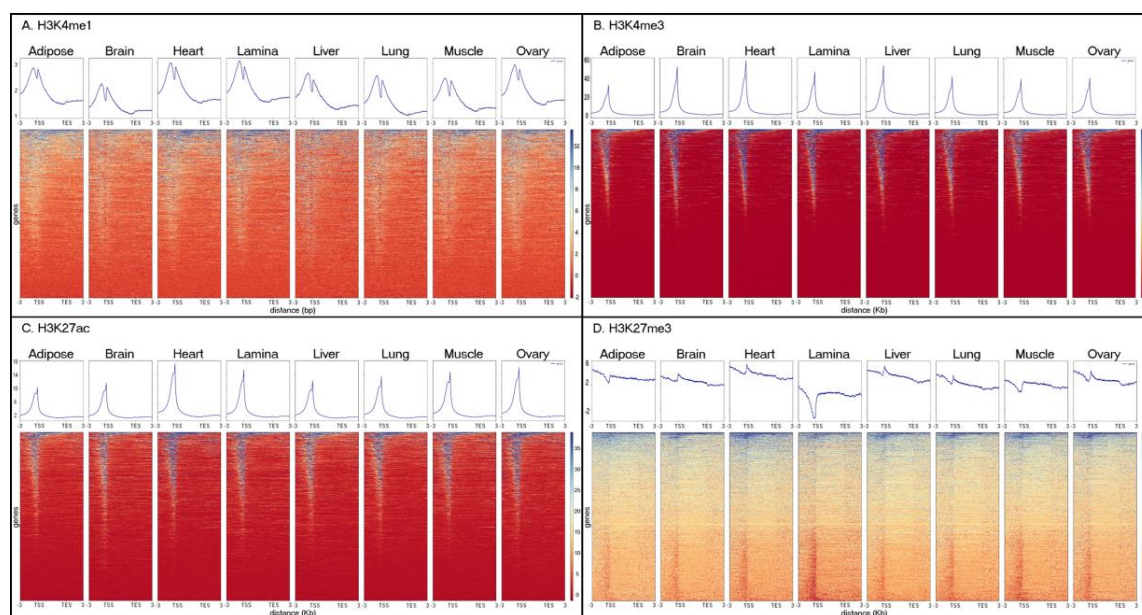


Figure 6. Histone mark enrichment across the average annotated gene body. (A) H3K4me1, (B) H3K4me3, (C) H3K27ac, (D) H3K27me3 from SICER. Topology plots (top) and heat maps (bottom) show average enrichment of each histone mark in each tissue across a size-normalized gene distribution based on the Ensembl (Release 95) annotation for EquCab3.0. Each line in the heatmap represents mark enrichment across a given gene, such that red indicates low relative enrichment and blue indicates high relative enrichment.

3.3. Identifying Motifs and Biological Process GO Terms

We identified between 16 and 61 transcription factor binding motifs in the unique active regions for each tissue, and the full results from the GO term analysis can be found in Tables S4–S19. While a large proportion of the identified transcription factors were still shared between tissues despite being identified in tissue-specific active regions, all of the tissues except for adipose and lung had at least one uniquely detected transcription factor binding site. A uniquely detected binding site was defined as a motif that was only identified in the tissue-specific active regions for a single tissue. The motifs were ranked based on significance of enrichment after multiple testing correction, and the top five enriched and identified motifs for each tissue are listed in Table 5. Ovary was the only tissue that had uniquely detected transcription factor motifs in this top five list, and the most significant motif was associated with FOXO3, which is a transcription factor (TF) characterized by 39 GO terms, including several for tissue-specific functions such as oocyte maturation (GO:0001556) and ovulation (GO:0001542). In fact, upon closer inspection, the FOXO3 motif was found within a tissue-specific regulatory element near a gene with recognized roles in mammalian reproduction, *NR5A1* (Figure 7) [52]. Using a network analysis for TFs implicated in each tissue, we found that six networks contained EP300 as a central node, although these networks did not link every TF for a given tissue (Figure 8A). Interestingly, MYC was the central node for brain, liver, and skeletal muscle (Figure 8B), and TP53 was the central node for lung (Figure 8C).

Table 5. Top five significantly enriched transcription factor binding motifs identified in tissue-specific active enhancers. Tissue specificity of the active enhancers was defined by overlap of H3K4me1 and H3K27ac in the same tissue and no overlap of these marks in this region in any other tissues, and tissue specificity of the binding motifs was then defined by detection of an enriched binding motif in only one tissue.

Rank	Motif ID	Consensus	Adjusted <i>p</i> -Value	UniProt Entry
Adipose				
1	SP3	VCCACGCCCMC	1.49×10^{-10}	Q02447
2	TFDP1	VSGCGGGA AVN	1.74×10^{-10}	Q14186
3	TFAP2A	HGCCYSAGGCD	3.27×10^{-10}	P05549
4	TFAP2C	YGCCYBVRGGCA	4.56×10^{-10}	Q92754
6	KLF16	GMCACGCCCCC	5.81×10^{-9}	Q9BXX1
Brain				
1	TFAP2A(var.2)	YGCCCBVRGGCR	1.82×10^{-16}	P05549
2	TFAP2B	YGCCCBVRGGCA	1.29×10^{-13}	Q92481
3	SP3	VCCACGCCCMC	2.69×10^{-13}	Q02447
4	TFAP2C	YGCCYBVRGGCA	4.99×10^{-13}	Q92754
5	KLF16	GMCACGCCCCC	1.06×10^{-12}	Q9BXX1
Heart				
1	MZF1	BGGGGA	2.23×10^{-5}	P28698
2	Ascl2	ARCAGCTGCT	7.06×10^{-4}	Q99929
3	ASCL1	VSAGCAGCTGSNN	9.41×10^{-4}	P50553
4	SP3	VCCACGCCCMC	1.42×10^{-3}	Q02447
5	NEUROD1	NRACAGATGGYNN	1.60×10^{-3}	Q13562
Lamina				
1	SP2	GYCCCGCCYCYBSSS	8.51×10^{-15}	Q02086
2	SP1	GCCCCKCCCCC	5.98×10^{-14}	P08047
3	SP3	VCCACGCCCMC	2.84×10^{-13}	Q02447
4	KLF16	GMCACGCCCCC	3.22×10^{-13}	Q9BXX1
7	Zfx	SSSGCCBVGCCCTS	1.06×10^{-11}	P17010
Liver				
1	SP1	GCCCCKCCCCC	7.81×10^{-13}	P08047
2	TFAP2B	YGCCCBVRGGCA	4.23×10^{-12}	Q92481
3	TFAP2C	YGCCYBVRGGCA	1.87×10^{-11}	Q92754
4	TFAP2A	HGCCYSAGGCD	4.97×10^{-11}	P05549
5	ZNF740	MCCCCCCCAC	8.99×10^{-11}	Q8NDX6
Lung				
1	THAP1	YTGCCCDDBA	5.09×10^{-9}	Q9NVV9
3	ESR2	AGGTCASVNTGMCCY	1.08×10^{-8}	Q92731
4	Zfx	SSSGCCBVGCCCTS	1.44×10^{-8}	P17010
5	ZBTB7A	NVCCGGAAGTGSV	1.46×10^{-8}	O95365
6	TFAP2A(var.2)	YGCCCBVRGGCR	6.51×10^{-8}	P05549
Muscle				
1	SP1	GCCCCKCCCCC	1.45×10^{-9}	P08047
2	SP2	GYCCCGCCYCYBSSS	4.21×10^{-9}	Q02086
3	SP8	RCCACGCCCMCY	1.15×10^{-8}	Q8IXZ3
4	CTCF	CRSCAGGGGGCRSB	3.44×10^{-8}	Q8NI51
5	KLF16	GMCACGCCCCC	4.36×10^{-7}	Q9BXX1
Ovary				
1	FOXO3 ¹	DAAAYA	7.23×10^{-7}	O43524
3	KLF16	GMCACGCCCCC	1.93×10^{-4}	Q9BXX1
4	FOXC1 ¹	WAWGTAAAYAW	2.39×10^{-4}	Q12948
6	CTCF	CRSCAGGGGGCRSB	4.38×10^{-4}	Q8NI51
7	Arid5a	SYAATATTGVDANH	4.99×10^{-4}	Q03989

¹ Transcription factor (TF) motifs that were only detected in one tissue.

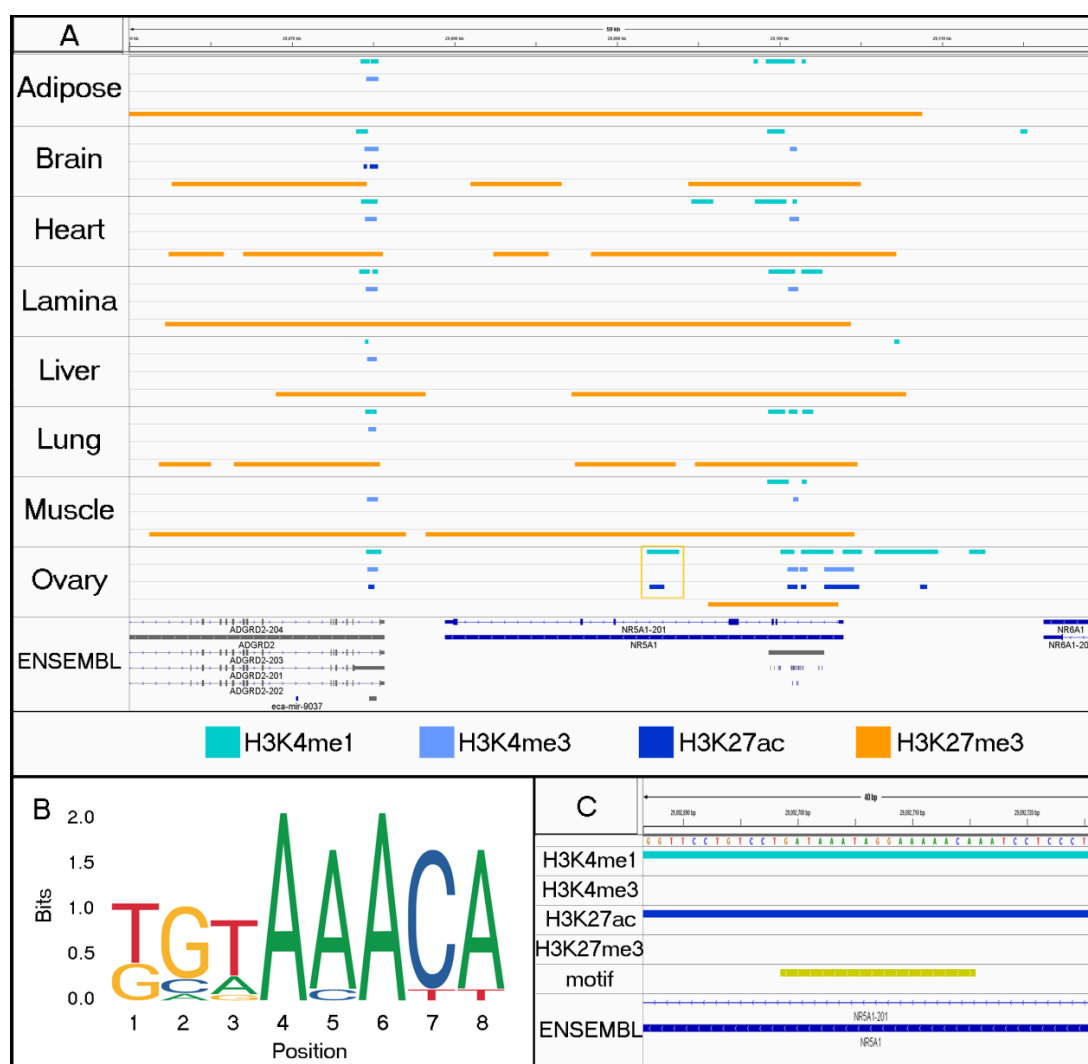


Figure 7. Localizing enriched TF binding motifs within tissue-specific peaks. For each tissue, peaks are displayed for H3K4me1 (cyan), H3K4me3 (light blue), H3K27ac (dark blue), and H3K27me3 from SICER (orange). Gold box highlights ovary specific marks in intron of *NR5A1*. **(B)** Motif logo displays one of the major motifs for FOXO3. **(C)** Zoomed in view of 40 bp within region highlighted in **(A)**. The gold track indicates the presence of two nearly consecutive motifs for FOXO3 within the H3K4me1 and H3K27ac peaks that were detected only in ovary tissue.

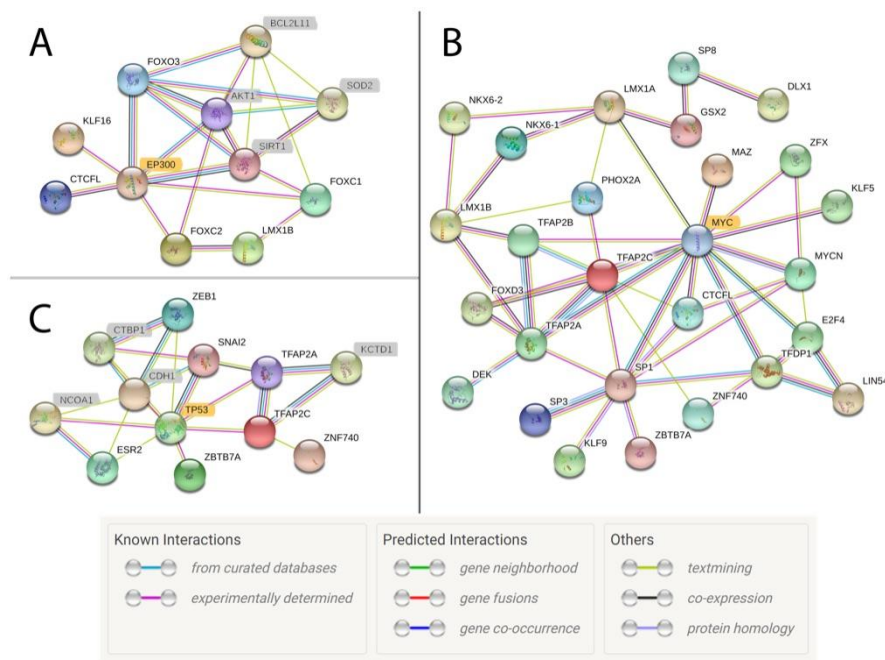


Figure 8. Network of transcription factor interactions based on active motif regions in ovary, skeletal muscle, and lung tissues. Each circle represents a protein, and lines indicate interactions between proteins. Each line color indicates a different type of evidence for the interaction. Protein names highlighted in grey were not identified in the horse ChIP-seq data but were identified as secondary factors by the network analysis based on implicated or known protein–protein interactions from research in humans and mice. (A) EP300 is highlighted in gold to denote its central role in the network for ovary tissue. (B) MYC is highlighted in gold to indicate its central role in the network for muscle tissue. (C) TP53 is highlighted in gold to indicate its central role in the network for lung. Legend is from String Software [47].

4. Discussion

As part of the FAANG consortium, we mapped more than 1 million putative regulatory sites across the equine genome, which will contribute significantly to our understanding of genome regulation in horses, as well as regulatory differences across species. To ensure that we obtained high quality data for the horse, we first performed QC experiments to optimize tissue-specific laboratory protocols (Table S2). Adipose tissue presented a distinct challenge for chromatin extraction due to the low number of nuclei and large amount of cellular material, including lipid deposits. Efforts to obtain high quality data were successful, as quality metrics for this tissue surpassed ENCODE standards or scored within the range of the other tissues for a given mark. From this work, we suspect that other difficult tissues (i.e., those with low nuclei density, extensive and persistent extra- or intracellular material, etc.) will also require similar alternate approaches for chromatin extraction, such as additional starting material or specialized kits.

To ensure accurate and relevant peak-calling, we employed MACS2 software for calling H3K4me1, H3K4me3, and H3K27ac peaks and both MACS2 and SICER for H3K27me3 peaks. While several attempts have been made to develop a bench-marking method for ranking ChIP peak-callers, there is no established gold standard for selecting a particular application [53]. Steinhauser et al., for example, found that SICER was better able to detect true-positive differential regions (DRs) when H3K36me3 data were down-sampled to as low as 10%, while MACS2 was only effective at detecting true-positive DRs at 60% or higher [54]. SICER, however, had a higher number of false-positive DRs compared to MACS2, indicating that it may be sacrificing some specificity to obtain higher sensitivity, while MACS2 does the opposite. Based upon generating peaks with higher identity between replicates (Figure S1) and higher proportions of the genome covered (Table 3), we found SICER software to be

more consistent for calling broad peak-calls compared to MACS2, which is consistent with previous research [55]. For that reason, we continued our investigations using the SICER-called H3K27me3 peaks. Given that peaks from both callers are supported in the literature for other species studied, both sets of peak-calls are available at <https://data.faang.org/dataset/PRJEB35307>.

Using the Jaccard Similarity Index to compare significant peak-calls, we determined that biological replicates were highly similar for each tissue with the exception of the brain replicates for H3K4me1 and the ovary replicates for H3K27me3 (Figure 1). While these low identity scores could be the result of underlying biological differences between the samples of brain or ovary tissue, this is unlikely given that the differences between replicates were only found in one mark for each of the two tissues. Using a bioinformatic method that relied on identifying overlapping enrichment rather than the intersection of significant peaks ensured accurate combined peak calling despite lower similarity scores. Combined peak-calls were assessed for significant enrichment surrounding genes with known expression patterns. In particular, the three active marks (H3K4me1, H3K4me3, and H3K27ac) had consistent peaks between tissues near *ACTB* (Figure 2A) and tissue-specific peaks near the liver enzyme *CYP2E1* (Figure 2B), which support the validity of our peak-calling methods. Additionally, these peak-calls were all consistent with RNA expression from the same tissues for ECA_UCD_AH1 and ECA_UCD_AH2, further supporting the validity of the enrichment patterns detected for each histone mark.

Comparing replicated peaks between tissues, we found that liver, brain, and lamina had the highest percentage of peaks that were unique to only that tissue (Figure 3). Liver is known to have hundreds of distinct biological functions [48]. Since many of the functions are entirely unique to liver, a high proportion of unique active regulatory elements is consistent with the specialization of this tissue. Additionally, the mammalian brain is thought to have hundreds of different neuronal cell types within the cortex to coordinate numerous neurologic functions simultaneously, making it one of the most complex tissues of the body [56]. Therefore, these results are consistent with the expectation that there is a high degree of regulatory specificity to control numerous coordinated functions within complex tissues.

Interestingly, lamina was the other tissue with a high percentage of unique peaks for H3K4me1, H3K4me3, and H3K27ac (Figure 3). Due to the role of lamina in disease, this finding may be particularly impactful for research into the physiological changes associated with laminitis, a syndrome that was established as a priority for equine research by the American Association of Equine Practitioners [57,58]. Additionally, when comparing the distribution of marks across the average gene body, we found that lamina tissue had an unusually low level of the repressive mark, H3K27me3, at the TSS (Figure 6). While all of the tissues had a portion of genes that appear to have a dip in enrichment directly at the TSS, lamina had the most extreme decrease. Perhaps the dip in enrichment for lamina is an indication of increased levels of expression across more of the genome and, when combined with the high percentage of unique peaks, could suggest that hoof lamina may perform additional biological processes that are currently uncharacterized. In order to understand all of the molecular functions of lamina tissue and its role in laminitis, more work is needed to further annotate and functionally characterize the extent of the cellular processes within healthy and diseased tissue. Without further validation, however, we cannot exclude that this decrease in enrichment may be a technical artifact such as decreased detection of H3K27me3 in lamina tissue due to low cell numbers and excess extracellular matrix.

To further characterize tissue-specific regulatory regions, we identified numerous transcription factor binding motifs within the unique active elements for each tissue (Table 5). Despite investigating elements that were only enriched in one tissue, we still identified many transcription factor motifs that were shared between tissues (Table 4 and Table S19) and, based on a GO term analysis, many of these factors had numerous associated biological processes (Tables S4–S18). Our network analysis identified EP300 as a central secondary factor needed to connect many implicated TFs (Figure 8A). This gene encodes a histone acetyltransferase that is known to interact with many TFs by protein–protein interactions rather than DNA binding [59,60]. Given that the TF motifs were all identified within active elements based on the presence of H3K27ac, finding a connection with EP300 supports our ability to

detect relevant peaks for this histone modification. Interestingly, there were three tissues (brain, liver, and muscle) that had MYC rather than EP300 as the central node for their TF networks (Figure 8B). We found that binding motifs for MYC were enriched in all three tissues, which is consistent with its role as a TF for many common cellular processes, including growth and regulation of the cell cycle [61]. Similarly, TP53 was the central node in the protein interaction network for lung tissue (Figure 8C), which is consistent with its role as a key regulator of the cell cycle when functional and as a major tumorigenesis factor in lung cancer when dysfunctional [62].

By developing tissue-specific maps of the H3K4me1, H3K4me3, H3K27ac, and H3K27me3 histone modifications, we aimed to build upon our current knowledge about genomic regulation in the horse and provide new resources for further advancing research on health, performance, and reproduction traits. In particular, incorporating histone ChIP-seq data with current annotations is expected to lead to the discovery of more tissue-specific functional variants that are outside of protein-coding regions. For example, a genome-wide investigation of liver fibrosis in the Swiss Franches-Montagnes breed led to the identification of an associated region containing a promising candidate gene, *PKHD1*, which was also associated with kidney and liver disease in humans [49]. Sequencing the coding region of this gene identified two synonymous coding variants strongly associated with disease; however, their causal role remains unclear. Comparing peaks from multiple tissues in that region, there are also several liver-specific regulatory peaks within intron 59 of *PKHD1* (Ensembl Transcript ID: ENSECAT00000024985.1) that may affect expression of this or another hepatic gene nearby (Figure 4) and represent another avenue to explore for the molecular mechanism underlying liver fibrosis.

Tissue-specific histone modifications have also been implicated in the complex traits of performance [63] and, throughout modern horse domestication, performance abilities such as speed have been a major factor for selective breeding [64]. For example, recent research comparing three trotting breeds in Scandinavia (North Swedish Draught, Norwegian–Swedish Coldblooded trotter, and Standardbred trotter) found a 684 Kb genomic region associated with trotting racing ability containing numerous annotated and unannotated genes [50]. Using a protein-coding annotation, only one gene in the region, *TRIM37*, was identified as a potential candidate due to its associations with growth phenotypes in humans. Utilizing the locations of histone mark enrichment could help to identify other candidates such as tissue-specific regulatory regions or the genes that they regulate. In particular, there are H3K4me3 peaks at the start of a novel lncRNA predicted by Ensembl genebuild [51] within the region of interest for racing ability (Figure 5A). While uncharacterized, *LOC111775680* has a set of unique promoters found in skeletal and cardiac muscle with corresponding transcript data (as determined by RNAseq of the same tissues from ECA_UCD_AH1 and ECA_UCD_AH2), suggesting that the lncRNA may play a role in muscle physiology (Figure 5B). Additional work is needed to further characterize *LOC111775680* and investigate its potential role in muscle tissue and any corresponding effects on racing ability.

In addition to health and performance, these annotation data can also be used to identify important genomic regulatory regions for the reproductive system. Previously, homology with humans and mice was utilized to create a panel of candidate regions for determining the genetic cause of gonadal dysgenesis disorders across many breeds of horses [65], yet the panel was focused on coding variants and those within untranslated regions that may affect RNA and protein synthesis. By evaluating histone marks, we found that there are additional regulatory regions relevant to the reproductive system that were missed by the homology-based approaches. For example, *NR5A1*, a gene that is implicated in many dysgenesis cases [52,66], has a regulatory region as identified by H3K4me1 and H3K27ac histone modifications in equine ovary tissue that is characteristic of an active enhancer (Figure 7A). Upon further inspection, we identified two FOXO3 binding sites within the region enriched with peak-calls (Figure 7B,C). FOXO3 is a transcription factor with known roles in oocyte maturation [GO:0001556] and ovulation [GO:0001542] that was identified as an ovary-specific TF based on our motif analysis (Table S18).

Along with the regulatory regions identified in this study, the equine FAANG consortium is characterizing the full RNA profile of more than 50 tissues and cell types from the FAANG equine biobank supported by a research community initiative [26] and the DNA methylation profiles for the TOI. Moreover, the equine FAANG consortium is currently mapping a major insulator protein known as CTCF, with the goal of generating tissue-specific chromatin state predictions for the eight TOI. CTCF plays a key role in defining chromatin looping and, therefore, topologically associated domains and gene-enhancer interactions when combined with histone ChIP-Seq [67,68]. The same panel of genomic investigations are also being conducted in two stallions, so that males and females are represented in the annotations. We anticipate that the integration and utilization of these functional annotation datasets by the equine genomics community will lead to the identification of causal, non-coding variants underlying many traits of interest for equine medicine, performance, and reproduction.

Supplementary Materials: The following are available online at <http://www.mdpi.com/2073-4425/11/1/3/s1>: Figure S1. Jaccard Index to measure similarity of peaks called for H3K27me3 peaks based on either MACS2 or SICER peak-calling; Table S1. Concentrations and catalog numbers for antibodies used during ChIP; Table S2. Quality metrics, peak-calling summary, and comparison to Encyclopedia of DNA Elements (ENCODE) data from quality control (QC) testing; Table S3. Comparing peak numbers to ENCODE data; Table S4. Motif and gene ontology (GO) term analysis results for adipose-specific active enhancers; Table S5. Motif and GO term analysis results for adipose-specific active promoters; Table S6. Motif and GO term analysis results for brain-specific active enhancers; Table S7. Motif and GO term analysis results for brain-specific active promoters; Table S8. Motif and GO term analysis results for heart-specific active enhancers; Table S9. Motif and GO term analysis results for heart-specific active promoters; Table S10. Motif and GO term analysis results for lamina-specific active enhancers; Table S11. Motif and GO term analysis results for lamina-specific active promoters; Table S12. Motif and GO term analysis results for liver-specific active enhancers; Table S13. Motif and GO term analysis results for liver-specific active promoters; Table S14. Motif and GO term analysis results for lung-specific active enhancers; Table S15. Motif and GO term analysis results for lung-specific active promoters; Table S16. Motif and GO term analysis results for muscle-specific active enhancers; Table S17. Motif and GO term analysis results for muscle-specific active promoters; Table S18. Motif and GO term analysis results for ovary-specific active enhancers; Table S19. Summary of the GO term analysis results for unique active elements from all tissues.

Author Contributions: Conceptualization, C.J.F., R.R.B., J.L.P., T.S.K., and J.N.M.; methodology, C.C., C.K., R.R.B., C.J.F., and J.L.P.; software, N.B.K. and C.K.; formal analysis, N.B.K.; investigation, N.B.K. and C.C.; resources, C.J.F., R.R.B., J.L.P., E.N.H., and H.Z.; data curation, N.B.K.; writing—original draft preparation, N.B.K. and R.R.B.; writing—review and editing, C.J.F., J.L.P., J.N.M., T.S.K., C.K., C.C., E.N.H., and H.Z.; visualization, N.B.K.; supervision, R.R.B.; project administration, C.J.F., R.R.B., and J.L.P.; funding acquisition, C.J.F., R.R.B., J.L.P., T.S.K., and J.N.M. All authors have read and agreed to the published version of the manuscript.

Funding: Funding for experimental materials, and other resources was provided by the Grayson Jockey Club Foundation, USDA NRSP-8 equine species coordinator funds, and the Center for Equine Health with funds provided by the State of California pari-mutuel fund and contributions by private donors. Support for N.B.K. was also provided by the Grayson Jockey Club Foundation, Morris Animal Foundation (D16-EQ-028), and a fellowship from the Center for Equine Health at the University of California—Davis. Support for C.J.F. was provided by the National Institutes of Health (NIH) (K01OD015134 and L40 TR001136).

Acknowledgments: The authors would like to acknowledge members of the UC Davis FAANG group for providing their technical expertise and feedback. The authors would also like to acknowledge the contributions from horse owners that made this work possible.

Conflicts of Interest: C.C. is affiliated with Diagenode which is a leading provider of complete solutions for epigenetics research, including end-to-end services.

References

1. Rudolph, J.A.; Spier, S.J.; Byrns, G.; Hoffman, E.P. Linkage of hyperkalaemic periodic paralysis in Quarter horses to the horse adult skeletal muscle sodium channel gene. *Anim. Genet.* **1992**, *23*, 241–250. [\[CrossRef\]](#)
2. Finno, C.J.; Bannasch, D.L. Applied equine genetics. *Equine Vet. J.* **2014**, *46*, 538–544. [\[CrossRef\]](#)
3. Raudsepp, T.; Finno, C.J.; Bellone, R.R.; Petersen, J.J. Ten years of the horse reference genome: Insights into equine biology, domestication, and population dynamics in the post-genome era. *Anim. Genet.* **2019**, *50*, 569–597. [\[CrossRef\]](#)
4. OMIA—Online Mendelian Inheritance in Animals. Available online: <https://omia.org/home/> (accessed on 4 January 2019).

5. Wade, C.M.; Giulotto, E.; Sigurdsson, S.; Zoli, M.; Gnerre, S.; Imsland, F.; Lear, T.L.; Adelson, D.L.; Bailey, E.; Bellone, R.R.; et al. Genome Sequence, Comparative Analysis, and Population Genetics of the Domestic Horse. *Science* **2009**, *326*, 865–867. [[CrossRef](#)]
6. Mathelier, A.; Shi, W.; Wasserman, W.W. Identification of altered cis-regulatory elements in human disease. *Trends Genet.* **2015**, *31*, 67–76. [[CrossRef](#)]
7. Koufariotis, L.; Chen, Y.P.; Bolormaa, S.; Hayes, B.J. Regulatory and coding genome regions are enriched for trait associated variants in dairy and beef cattle. *BMC Genom.* **2014**, *15*, 436. [[CrossRef](#)]
8. Wang, M.; Hancock, T.P.; MacLeod, I.M.; Pryce, J.E.; Cocks, B.G.; Hayes, B.J. Putative enhancer sites in the bovine genome are enriched with variants affecting complex traits. *Genet. Sel. Evol.* **2017**, *49*, 56. [[CrossRef](#)]
9. Danko, C.G.; Choate, L.A.; Marks, B.A.; Rice, E.J.; Wang, Z.; Chu, T.; Martins, A.L.; Dukler, N.; Coonrod, S.A.; Tait Wojno, E.D.; et al. Dynamic evolution of regulatory element ensembles in primate CD4⁺ T cells. *Nat. Ecol. Evol.* **2018**, *2*, 537–548. [[CrossRef](#)] [[PubMed](#)]
10. Pennacchio, L.A.; Bickmore, W.; Dean, A.; Nobrega, M.A.; Bejerano, G. Enhancers: Five essential questions. *Nat. Rev. Genet.* **2013**, *14*, 288–295. [[CrossRef](#)] [[PubMed](#)]
11. Pasquali, L.; Gaulton, K.J.; Rodriguez-Segui, S.A.; Mularoni, L.; Miguel-Escalada, I.; Akerman, I.; Tena, J.J.; Moran, I.; Gomez-Marin, C.; van de Bunt, M.; et al. Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat. Genet.* **2014**, *46*, 136–143. [[CrossRef](#)] [[PubMed](#)]
12. Sun, W.; Poschmann, J.; del Rosario, R.C.H.; Parikshak, N.N.; Hajan, H.S.; Kumar, V.; Ramasamy, R.; Belgard, T.G.; Elangovan, B.; Wong, C.C.; et al. Histone Acetylome-wide Association Study of Autism Spectrum Disorder. *Cell* **2016**, *167*, 1385–1397. [[CrossRef](#)] [[PubMed](#)]
13. The FAANG Consortium; Andersson, L.; Archibald, A.L.; Bottema, C.D.; Brauning, R.; Burgess, S.C.; Burt, D.W.; Casas, E.; Cheng, H.H.; Clarke, L.; et al. Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome Biol.* **2015**, *16*, 57. [[CrossRef](#)] [[PubMed](#)]
14. Tuggle, C.K.; Guiffra, E.; White, S.N.; Clarke, L.; Zhou, H.; Ross, P.J.; Acloque, H.; Reecy, J.M.; Archibald, A.; Bellone, R.R.; et al. GO-FAANG meeting: A Gathering On Functional Annotation of Animal Genomes. *Anim. Genet.* **2016**, *47*, 528–533. [[CrossRef](#)] [[PubMed](#)]
15. Guiffra, E.; Tuggle, C.K.; FAANG Consortium. Functional Annotation of Animal Genomes (FAANG): Current Achievements and Roadmap. *Annu. Rev. Anim. Biosci.* **2019**, *7*, 65–88. [[CrossRef](#)]
16. Bannister, A.J.; Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Res.* **2011**, *21*, 381–395. [[CrossRef](#)]
17. Allfrey, V.G.; Faulkner, R.; Mirsky, A.E. Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis. *Proc. Natl. Acad. Sci. USA* **1964**, *51*, 786–794. [[CrossRef](#)]
18. Howe, F.S.; Fischl, H.; Murray, S.C.; Mellor, J. Is H3K4me3 instructive for transcription activation? *Bioessays* **2016**, *39*, 1–12. [[CrossRef](#)]
19. Calo, E.; Wysocka, J. Modification of enhancer chromatin: What, how, and why? *Mol. Cell* **2013**, *49*, 825–837. [[CrossRef](#)]
20. Entrevan, M.; Schuettengruber, B.; Cavalli, G. Regulation of Genome Architecture and Function by Polycomb Proteins. *Trends Cell Biol.* **2016**, *26*, 511–525. [[CrossRef](#)]
21. Tie, F.; Banerjee, R.; Stratton, C.A.; Prasad-Sinha, J.; Stepanik, V.; Zlobin, A.; Diaz, M.O.; Scacheri, P.C.; Harte, P.J. CBP-mediated acetylation of histone H3 lysine 27 antagonizes Drosophila Polycomb silencing. *Development* **2009**, *136*, 3131–3141. [[CrossRef](#)]
22. Spange, S.; Wagner, T.; Heinzel, T.; Kramer, O.H. Acetylation of non-histone proteins modulates cellular signaling at multiple levels. *Int. J. Biochem. Cell Biol.* **2009**, *41*, 185–198. [[CrossRef](#)] [[PubMed](#)]
23. Creighton, M.P.; Cheng, A.W.; Welstead, G.G.; Kooistra, T.; Carey, B.W.; Steine, E.J.; Hanna, J.; Lodato, M.A.; Frampton, G.M.; Sharp, P.A.; et al. Histone H3K27ac separates active from poised enhancers and predicts developmental states. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 21931–21936. [[CrossRef](#)] [[PubMed](#)]
24. Bernstein, B.E.; Humphrey, E.L.; Erlich, R.L.; Schneider, R.; Bouman, P.; Liu, J.S.; Kouzarides, T.; Schreiber, S.L. Methylation of histone H3 Lys 4 in coding regions of active genes. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 8695–8700. [[CrossRef](#)] [[PubMed](#)]
25. Dai, H.; Wang, Z. Histone Modification Patterns and Their Responses to Environment. *Curr. Environ. Health Rep.* **2014**, *1*, 11–21. [[CrossRef](#)]

26. Burns, E.N.; Bordbari, M.H.; Mienaltowski, M.J.; Affolter, V.K.; Barro, M.V.; Gianino, F.; Gianino, G.; Giulotto, E.; Kalbfleisch, T.S.; Katzman, S.A.; et al. Generation of an equine biobank to be used for Functional Annotation of Animal Genomes project. *Anim. Genet.* **2018**, *49*, 564–570. [\[CrossRef\]](#)
27. Landt, S.G.; Marinov, G.K.; Kundaje, A.; Kheradpour, P.; Pauli, F.; Batzoglou, S.; Bernstein, B.E.; Bickel, P.; Brown, J.B.; Cayting, P.; et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **2012**, *22*, 1813–1831. [\[CrossRef\]](#)
28. Ramirez, F.; Ryan, D.P.; Gruning, B.; Bhardwaj, V.; Kilpert, F.; Richter, A.S.; Heyne, S.; Dundar, F.; Manke, T. deepTools2: A next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **2016**, *44*, W160–W165. [\[CrossRef\]](#)
29. Trim Galore. Version 0.4.0. Available online: http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.
30. Martin, M. Cutadapt Removes Adapter Sequences From High-Throughput Sequencing Reads. *EMBnet J.* **2011**, *17*, 10–12. [\[CrossRef\]](#)
31. Kalbfleisch, T.S.; Rice, E.S.; DePriest, M.S.; Walenz, B.P.; Hestand, M.S.; Vermeesch, J.R.; O’Connell, B.L.; Fiddes, I.T.; Vershinina, A.O.; Saremi, N.F.; et al. Improved reference genome for the domestic horse increases assembly contiguity and composition. *Commun. Biol.* **2018**, *1*, 197. [\[CrossRef\]](#)
32. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760. [\[CrossRef\]](#)
33. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [\[CrossRef\]](#)
34. Picard Toolkit, GitHub Repository. Version 2.7.1. Available online: <http://broadinstitute.github.io/picard/>.
35. Zhang, Y.; Liu, T.; Meyer, C.A.; Eickhout, J.; Johnson, D.S.; Bernstein, B.E.; Nusbaum, C.; Myers, R.M.; Brown, M.; Wei, L.; et al. Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **2008**, *9*, R137. [\[CrossRef\]](#)
36. SICERpy, GitHub Repository. Available online: <https://github.com/dariober/SICERpy> (accessed on 10 October 2019).
37. Zang, C.; Schones, D.E.; Zeng, C.; Cui, K.; Zhao, K.; Peng, W. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics* **2009**, *25*, 1952–1958. [\[CrossRef\]](#) [\[PubMed\]](#)
38. Favorov, A.; Mularoni, L.; Cope, L.M.; Medvedeva, Y.; Mironov, A.A.; Makeev, V.J.; Wheelan, S.J. Exploring Massive, Genome Scale Datasets with the GenometriCorr Package. *PLoS Comput. Biol.* **2012**, *8*, e1002529. [\[CrossRef\]](#) [\[PubMed\]](#)
39. Quinlan, A.R.; Hall, I.M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **2010**, *26*, 841–842. [\[CrossRef\]](#)
40. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*; Springer: New York, NY, USA, 2016.
41. R Core Team. *R: A Language and Environment for Statistical Computing*; Version 3.4.3; R Foundation for Statistical Computing: Vienna, Austria, 2018. Available online: <https://www.r-project.org/>.
42. McLeary, R.; Bailey, T.L. Motif Enrichment Analysis: A unified framework and method evaluation. *BMC Bioinform.* **2010**, *11*, 165.
43. Khan, A.; Fornes, O.; Stigliani, A.; Gheorghe, M.; Castro-Mondragon, J.A.; van der Lee, R.; Bessy, A.; Cheneby, J.; Kulkarni, S.R.; Tan, G.; et al. JASPAR 2018: Update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.* **2018**, *46*, D260–D266. [\[CrossRef\]](#)
44. The UniProt Consortium. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* **2019**, *47*, D506–D515. [\[CrossRef\]](#)
45. Robinson, J.T.; Thorvaldsdottir, H.; Winckler, W.; Guttman, M.; Lander, E.S.; Getz, G.; Mesirov, J.P. Integrative Genomics Viewer. *Nature Biotechnol.* **2011**, *29*, 24–26. [\[CrossRef\]](#)
46. Aken, B.L.; Ayling, S.; Barrel, D.; Clarke, L.; Curwen, V.; Fairley, S.; Fernandez Banet, J.; Billis, K.; Garcia Giron, C.; Hourlier, T.; et al. The Ensembl gene annotation system. *Database* **2016**, *2016*, baw093. [\[CrossRef\]](#)
47. Szklarczyk, D.; Gable, A.L.; Lyon, D.; Junge, A.; Wyder, S.; Huerta-Cepas, J.; Simonovic, M.; Doncheva, N.T.; Morris, J.H.; Bork, P.; et al. STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **2019**, *47*, D607–D613. [\[CrossRef\]](#) [\[PubMed\]](#)
48. Trefts, E.; Gannon, M.; Wasserman, D.H. The liver. *Curr. Biol.* **2017**, *27*, R1141–R1155. [\[CrossRef\]](#) [\[PubMed\]](#)

49. Drogemuller, M.; Jagannathan, V.; Welle, M.M.; Graubner, C.; Straub, R.; Gerber, V.; Burger, D.; Signer-Hasler, H.; Poncet, P.A.; Klopfenstein, S. Congenital hepatic fibrosis in the Franches-Montagnes horse is associated with the polycystic kidney and hepatic disease 1 (*PKHD1*) gene. *PLoS ONE* **2014**, *9*, e110125. [[CrossRef](#)] [[PubMed](#)]
50. Velie, B.D.; Lillie, M.; Fegraeus, K.J.; Rosengren, M.K.; Sole, M.; Wiklund, M.; Ihler, C.F.; Strand, E.; Lindgren, G. Exploring the genetics of trotting racing ability in horses using a unique Nordic horse model. *BMC Genom.* **2019**, *20*, 104. [[CrossRef](#)] [[PubMed](#)]
51. Guttman, M.; Amit, I.; Garber, M.; French, C.; Lin, M.F.; Feldser, D.; Huarte, M.; Zuk, O.; Carey, B.W.; Cassady, J.P.; et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **2009**, *458*, 223–227. [[CrossRef](#)] [[PubMed](#)]
52. Ferraz-de-Souza, B.; Lin, L.; Achermann, J.C. Steroidogenic factor-1 (SF1, NR5A1) and human disease. *Mol. Cell. Endocrinol.* **2011**, *336*, 198–205. [[CrossRef](#)]
53. Szalkowski, A.M.; Schmid, C.D. Rapid innovation in ChIP-seq peak-calling algorithms is outdistancing benchmarking efforts. *Brief. Bioinform.* **2010**, *12*, 626–633. [[CrossRef](#)]
54. Steinhauser, S.; Kurzawa, N.; Eils, R.; Herrmann, C. A comprehensive comparison of tools for differential ChIP-seq analysis. *Brief. Bioinform.* **2016**, *17*, 953–966. [[CrossRef](#)]
55. Xu, S.; Grullon, S.; Ge, K.; Peng, W. Spatial Clustering for Identification of ChIP-Enriched Regions (SICER) to Map Regions of Histone Methylation Patterns in Embryonic Stem Cells. In *Stem Cell Transcriptional Networks*; Kidder, B., Ed.; Humana Press: New York, NY, USA, 2014; Volume 1150, pp. 97–111.
56. Masland, R.H. Neuronal cell types. *Curr. Biol.* **2004**, *14*, R497–R500. [[CrossRef](#)]
57. Belknap, J.K. Genomics of laminitis. In *Equine Genomics*, 1st ed.; Chowdhary, B.P., Ed.; John Wiley & Sons, Inc.: Ames, IA, USA, 2013; pp. 255–264.
58. Al-Agele, R.; Paul, E.; Kubale Dvojmc, V.; Sturrock, C.J.; Rauch, C.; Rutland, C.S. The Anatomy, Histology and Physiology of the Healthy and Lamé Equine Hoof. In *Veterinary Anatomy and Physiology*; Rutland, C., Kubale, V., Eds.; IntechOpen: London, UK, 2019; pp. 1–18.
59. Ogryzko, V.V.; Schiltz, R.L.; Russanova, V.; Howard, B.H.; Nakatani, Y. The Transcriptional Coactivators p300 and CBP Are Histone Acetyltransferases. *Cell* **1996**, *87*, 953–959. [[CrossRef](#)]
60. Song, C.; Keller, K.; Chen, Y.; Chen, Y.; Murata, K.; Stamatoyannopoulos, G. Transcription coactivator CBP has direct DNA binding activity and stimulates transcription factor DNA binding through small domains. *Biochem. Biophys. Res. Commun.* **2002**, *296*, 118–124. [[CrossRef](#)]
61. Oster, S.K.; Ho, C.S.; Soucie, E.L.; Penn, L.Z. The myc oncogene: Marvelously Complex. *Adv. Cancer Res.* **2002**, *84*, 81–154. [[PubMed](#)]
62. Mogi, A.; Kuwano, H. TP53 Mutations in Nonsmall Cell Lung Cancer. *J. Biomed. Biotechnol.* **2010**, *2011*, 9. [[CrossRef](#)] [[PubMed](#)]
63. Widmann, M.; Nieß, A.M.; Munz, B. Physical Exercise and Epigenetic Modifications in Skeletal Muscle. *Sports Med.* **2019**, *49*, 509–523. [[CrossRef](#)] [[PubMed](#)]
64. Fages, A.; Hanghoj, K.; Khan, N.; Gaunitz, C.; Seguin-Orlando, A.; Leonardi, M.; McCrory Constantz, C.; Gamba, C.; Al-Rasheid, K.A.S.; Albizuri, S.; et al. Tracking Five Millennia of Horse Management with Extensive Ancient Genome Time Series. *Cell* **2019**, *177*, 1419–1435. [[CrossRef](#)] [[PubMed](#)]
65. Pujar, S.; Meyers-Wallen, V.N. Sequence variations in equine candidate genes for XX and XY inherited disorders of sexual development. *Reprod. Domest. Anim.* **2012**, *47*, 827–834. [[CrossRef](#)]
66. Achermann, J.C.; Ozisik, G.; Ito, M.; Orun, U.A.; Harmanci, K.; Gurakan, B.; Jameson, J.L. Gonadal Determination and Adrenal Development Are Regulated by the Orphan Nuclear Receptor Steroidogenic Factor-1, in a Dose-Dependent Manner. *J. Clin. Endocrinol. Metab.* **2002**, *87*, 1829–1833. [[CrossRef](#)]
67. Splinter, E.; Heath, H.; Kooren, J.; Palstra, R.; Klous, P.; Grosveld, F.; Galiart, N.; de Laat, W.L. CTCF mediates long-range chromatin looping and local histone modification in the β -globin locus. *Genes Dev.* **2006**, *20*, 2349–2354. [[CrossRef](#)]
68. Hou, C.; Dale, R.; Dean, A. Cell type specificity of chromatin organization mediated by CTCF and cohesion. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 3651–3656. [[CrossRef](#)]

